Matrik: Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer
Vol. 25, No. 1, November 2025, pp. 39∼52
ISSN: 2476-9843, accredited by Kemenristekdikti, Decree No:10/C/C3/DT.05.00/2025
DOI: 10.30812/matrik.v25i1.5186

K-Means-Based Customer Segmentation with Domain-Specific Feature Engineering for Water Payment Arrears Management

Andi Hary Akbar[1], Heri Wijayanto[1], I Wayan Agus Arimbawa[1], Vynska Amalia Permadi[2]
[1]Universitas Mataram, Mataram, Indonesia
[2]University of Sheffield, Sheffield, United Kingdom

Article Info

Article history:

Received July 10, 2025
Revised October 13, 2025
Accepted October 25, 2025

Keywords:

Customer segmentation;
Feature engineering;
K-means clustering;
Payment behavior analysis;
Utility analytics.

ABSTRACT

Indonesian water utilities face persistent challenges in managing payment delinquencies due to diverse customer characteristics, geographic limitations, and inadequate analytical capabilities. Addressing this issue is essential to optimizing revenue collection and supporting sustainable operations. This study aims to develop a data-driven customer segmentation framework using K-means clustering to enhance delinquency management. The framework incorporates six engineered features—Debt Efficiency, Payment Behavior Score, Category Risk Score, Geographic Risk Score, Consumption Intensity, and Financial Risk Score—designed to capture customer payment behavior, consumption patterns, and geographic risk. We applied the model to 1,500 anonymized customer records from PT Air Minum Giri Menang, focusing on those with delinquencies exceeding four months. Risk scoring was based on quintile distribution, and optimal clustering was determined through the elbow method combined with silhouette coefficient analysis. The results produced a two-cluster solution (silhouette score = 0.538), showing statistically significant differences across features (p ¡ 0.001) and medium-to-large effect sizes (Cohen's d = 0.52–2.12). The segmentation identified medium-risk customers (86.7%) who require preventive management and high-risk customers (13.3%) who need billing intervention. Urban areas exhibited higher delinquency risk (18.4%) than rural areas (2.5%), indicating the need for geographically targeted strategies. All customer data was anonymized following Indonesian data protection protocols. In conclusion, the proposed framework transforms manual billing supervision into an adaptive, data-driven management system, contributing to segmentation research by introducing utility-specific engineered features for Indonesian water utilities.

Corresponding Author:

Andi Hary Akbar, +6285338838589,
Faculty of Engineering, Master of Information Technology,
Universitas Mataram, Mataram, Indonesia,
Email: i2s02410013@student.unram.ac.id.

How to Cite:
A. H. Akbar, H. Wijayanto, and I. W. A. Arimbawa, "K-Means-Based Customer Segmentation with Domain-Specific Feature Engineering for Water Payment Arrears Management", *MATRIK: Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer*, Vol. 25, No. 1, pp. 39-52, November, 2025.

1. INTRODUCTION

Customer segmentation provides an important analytical approach for modern organizations seeking to optimize revenue collection and enhance operational efficiency. The proliferation of data analytics across industries has enhanced how businesses understand customer behavior patterns. It has facilitated the deployment of machine learning-based segmentation techniques as powerful tools for identifying distinct customer groups and developing targeted management strategies [1, 2]. This analytical transformation is particularly crucial for water utility organizations, which face increasing challenges from diverse customer payment behaviors, geographic service constraints, and limited analytical capabilities for collection risk assessment, all of which threaten operational sustainability. Consequently, the adoption of advanced data analytics [3, 4], and the specific application of segmentation approaches have become important for utilities to maintain financial sustainability and operational effectiveness. These analytical capabilities enable utilities to move beyond simple demographic classification toward comprehensive behavioral analysis that predicts payment risks, optimizes collection strategies, and allocates resources more effectively across diverse customer populations.

Machine learning-based customer segmentation methodologies have demonstrated developments in recent years, with K-means clustering emerging as a particularly robust approach for practical business applications due to its computational efficiency and interpretable results. Li et al. [5] developed an enhanced K-means clustering framework that combines the algorithm with adaptive particle swarm optimization for customer segmentation, achieving improved clustering performance through systematic data preparation and algorithm optimization across multiple business contexts. Building on this algorithmic foundation, Rachman et al. [6] implemented Mini Batch K-means clustering specifically for credit card customer segmentation, demonstrating that computational efficiency improvements can be achieved while maintaining clustering quality for large-scale datasets. Similarly, Pradana [7] further optimized K-means clustering strategies for mall customer segmentation, achieving enhanced customer group identification through systematic hyperparameter optimization and comprehensive validation procedures that ensure clustering reliability. Complementing these algorithmic advances, comprehensive literature reviews have highlighted the broader methodological landscape and identified key research directions. Salminen et al. [8] systematically reviewed algorithmic customer segmentation approaches across various domains, identifying key methodological advances and establishing research directions for algorithm-based customer analysis that emphasize the growing importance of machine learning techniques. Concurrently, Alves Gomes and Meisen [9] conducted an extensive review of customer segmentation methods specifically for personalized targeting in e-commerce contexts, highlighting the effectiveness of clustering algorithms in identifying distinct behavioral patterns while emphasizing the critical role of feature engineering in achieving meaningful customer differentiation.

While these advances demonstrate clustering effectiveness in commercial settings, utility organizations face distinct challenges, including regulated pricing, essential service obligations, and geographic service constraints that require specialized segmentation approaches. Despite notable methodological developments in general customer segmentation, several gaps remain regarding specific applications within the utility sector. Current segmentation research predominantly focuses on commercial and e-commerce contexts [9], with limited examination of methodologies specifically designed to address payment collection challenges faced by water utility organizations. While comprehensive reviews [8] acknowledge algorithmic advances across domains, they highlight the lack of sector-specific feature engineering approaches. This sectoral gap creates several limitations that constrain practical utility applications. First, most existing studies rely on basic demographic and transactional features, often omitting domain-specific feature engineering that captures the unique payment behaviors, consumption characteristics, and operational risk factors inherent to utility customer populations [9]. Second, geographic considerations require domain-specific adaptation for utility contexts, as existing frameworks lack utility-specific geographic risk assessment methodologies despite their crucial role for operational planning [10]. Third, the integration of financial risk assessment methodologies with operational efficiency metrics has received insufficient research attention, a crucial oversight given the importance of comprehensive debt management for utility revenue optimization and sustainability [11]. These limitations highlight a significant gap in the segmentation literature: current methodologies lack comprehensive feature engineering that systematically integrates payment behavior analysis, consumption patterns, and geographic risk factors, thus impacting their adaptability in the utility sector.

This research addresses the identified gaps by developing a novel K-means clustering framework specifically designed for utility customer segmentation. It represents the first comprehensive integration of domain-specific feature engineering with geographic risk assessment methodologies for water utility contexts. Unlike previous studies that focus primarily on e-commerce contexts [9] or lack utility-specific feature engineering [8], this research innovates by systematically combining payment behavior analysis, consumption patterns, and territorial risk factors into six engineered features specifically designed for utility operations. K-means clustering was selected over alternative methods (hierarchical clustering, DBSCAN, Gaussian mixture models) based on four criteria essential for utility operational implementation: (1) interpretable cluster centroids enabling straightforward business communication to non-technical stakeholders, (2) computational efficiency supporting scalability to larger customer databases typical of regional water utilities, (3) proven effectiveness with continuous numerical features dominating our engineered feature set [12, 13],

and (4) deterministic reproducibility required for regulatory compliance. While more complex algorithms might capture non-linear relationships, K-means balances interpretability, efficiency, and operational deployment effectiveness, making it optimal for utility applications.

Our methodological approach builds upon a three-dimensional conceptual framework that integrates behavioral economics, operational risk management, and geographic accessibility theories. The framework posits that utility payment risk emerges from the interaction of three factors: (1) customer payment capability and willingness (captured through payment behavior and debt efficiency metrics), (2) service consumption patterns relative to financial obligations (reflected in consumption intensity and financial risk scores), and (3) operational accessibility constraints affecting collection effectiveness (quantified through geographic and category risk scores). This multidimensional conceptual model addresses limitations of single-dimension approaches by recognizing that payment behavior in utility contexts cannot be explained by demographics alone but requires an integrated assessment of financial behavior, consumption patterns, and operational context. The six engineered features operationalize this framework by transforming these theoretical dimensions into quantifiable risk metrics suitable for clustering analysis.

The primary purpose of this study is to develop a K-means clustering framework through domain-specific feature engineering that captures payment behaviors, consumption patterns, and geographic risk factors for effective arrears management in water utility contexts. We develop six engineered features that operationalize our conceptual framework: (1) Debt Efficiency normalizes outstanding amounts by consumption volume, enabling fair comparison across different customer usage levels; (2) Payment Behavior Score standardizes arrears duration into comparable risk scales; (3) Category Risk Score incorporates institutional knowledge about payment reliability patterns across customer types (residential, commercial, institutional); (4) Geographic Risk Score quantifies territorial accessibility and socioeconomic factors affecting collection difficulty; (5) Consumption Intensity uses logarithmic transformation to normalize usage patterns while preserving behavioral distinctions; and (6) Financial Risk Score provides composite assessment integrating debt magnitude, payment delays, and consumption efficiency. These features represent a departure from conventional demographic segmentation by focusing on behavioral and operational risk indicators directly relevant to optimizing collection management. Our methodological approach employs hyperparameter optimization through the combined application of the elbow method analysis and the silhouette coefficient evaluation to achieve improved clustering performance, complemented by comprehensive statistical validation procedures ensuring meaningful and actionable customer group differentiation.

This research addresses critical gaps in Indonesian water utility management by developing the first comprehensive machine learning-based customer segmentation framework specifically designed for the Indonesian drinking water sector. No previous studies have systematically applied advanced clustering methodologies to payment collection challenges. The implementation necessarily addresses personal data protection requirements established by Law No. 27 of 2022 [14] on Personal Data Protection (UU PDP) through anonymization protocols that balance analytical capabilities with privacy protection mandates for customer information processing. Applied to PTAM operational data from PTAM Giri Menang, this research makes contributions across multiple domains. For research science, we establish a methodological framework that advances customer segmentation literature by demonstrating domain-specific feature engineering adapted to local operational contexts and regulatory requirements. For community organizations and water utilities, we provide a practical, empirically validated framework that enables data-driven collection management, converting manual monitoring processes into systematic risk assessment tools. For government and regulatory bodies, our methodology offers evidence-based approaches for policy development in utility regulation that comply with national data protection standards. For the private sector, particularly utility companies and consulting firms, we demonstrate replicable methodologies that inform targeted payment collection strategies while maintaining regulatory compliance across diverse customer populations and geographic contexts.

2. RESEARCH METHOD

This research implements a K-means clustering methodology for utility customer segmentation. The methodology integrates feature engineering techniques with hyperparameter optimization to transform raw customer billing data into business intelligence. As illustrated in Figure 1, the research approach addresses three primary methodological objectives: first, transforming heterogeneous customer data into standardized risk assessment metrics through domain-specific feature engineering; second, applying K-means clustering algorithms with optimization to identify statistically different customer segments; and third, validating clustering results through statistical testing.

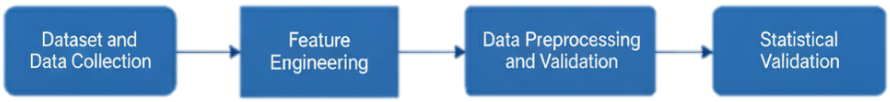

Figure 1. Research Methodology Framework for Utility Customer Segmentation

2.1. Dataset and Data Collection

The research utilizes customer billing and payment data from PT Air Minum Giri Menang (PTAM Giri Menang), a regional drinking water company serving Mataram and the West Lombok regions in Indonesia. All customer data underwent comprehensive anonymization procedures consistent with Indonesian Personal Data Protection Law (UU No. 27/2022) requirements. Customer identification information was replaced with anonymous identifiers, and geographic data was aggregated to the sub-district level to prevent individual identification. PTAM Giri Menang provided official institutional consent for the utilization of anonymous data for academic research purposes, ensuring all personal identifiers were removed before analysis, and data processing focused exclusively on aggregate payment behavior patterns.

The dataset represents customer records with payment arrears of four months or more, extracted from the company's operational billing system during the data collection period for this research. The four-month arrears threshold aligns with Indonesian water utility industry standards, which classify customers requiring collection intervention when there is a payment delay of 4+ months. This threshold represents both PTAM's standard operational trigger point for systematic collection activities and the industry-recognized boundary for analyzing customers requiring intervention strategies. It serves as the appropriate framework for collection management analysis rather than general customer segmentation. This study focuses specifically on customers with payment arrears, representing the subset of PTAM's customer base requiring collection intervention. Results apply primarily to collection management rather than general customer segmentation, as the sample excludes customers with current payment status. The geographic scope is limited to the PTAM Giri Menang service territory and requires validation before generalization to other Indonesian water utilities with different demographic compositions or service area characteristics.

The dataset contains 1,500 customer records with 11 attributes covering customer identification, demographic information, service classifications, payment arrears status, and water consumption data. The dataset structure includes customer identification information (Customer ID, Customer Name), geographic location details (Address, Village, Sub-district), service classifications (Status, Customer Category, Category Name), and financial and consumption metrics (Arrears Duration, Outstanding Amount, Water Consumption). All customer records in the dataset maintain an active service status. The geographic distribution spans 65 sub-districts and 121 villages, with customer concentrations in urban areas such as Cakranegara (12.3%), Sandubaya (10.1%), and Lembar (8.0%). The customer type distribution indicates a residential customer base, with Household Category B (56.0%) being the largest category, followed by Low-Income Housing B (20.5%) and Household Category A (15.8%). Commercial and institutional customers comprise smaller segments, including business categories (Micro Business, Small Business, Medium Business, Large Business) and government institutions (District/City Government Institutions, Provincial Government Institutions). The payment arrears data shows variation, with outstanding amounts ranging from IDR 144,200 to IDR 26,315,500 (mean = IDR 1,071,758, standard deviation = IDR 1,613,291), indicating heterogeneity in customer payment behavior and risk profiles.

2.2. Feature Engineering

Feature engineering transforms raw customer data into six engineered features designed to capture payment patterns, consumption behaviors, and territorial risk factors for effective segmentation analysis. Rather than relying solely on demographic characteristics, this approach develops behavioural and financial risk indicators that directly relate to payment collection patterns and operational requirements. The engineered features address the need for standardized risk assessment across diverse customer types while maintaining interpretability for practical business application [15]. The six engineered features are designed to capture different dimensions of customer risk behavior, as detailed in Table 1.

The development of these features addresses limitations in existing customer segmentation approaches for utility applications. Debt Efficiency addresses the limitation of absolute debt comparisons by normalizing outstanding amounts relative to service consumption. This results in values in IDR per cubic meter that range from thousands to hundreds of thousands, depending on payment efficiency. Traditional segmentation approaches compare raw debt values, which can disadvantage high-consumption customers who may have reasonable payment patterns relative to their usage. This metric enables comparison between different customer types, such as a residential customer with IDR 500,000 debt using 15 cubic meters versus a commercial customer with IDR 2,000,000 debt using 80 cubic meters. It shows that the residential customer has higher relative payment inefficiency despite the lower absolute debt amount.

The Category Risk Score incorporates domain-specific business knowledge that demographic segmentation often overlooks, generating risk weights ranging from 1 (lower risk categories) to 5 (higher risk categories). Category Risk Scores and Geographic Risk Scores are assigned automatically using quintile-based ranking [16] of debt efficiency metrics. Each customer category is ranked by its average debt efficiency (outstanding amount per cubic meter of consumption) and then divided into five quintiles, where quintile 1 receives a risk score of 1 (lowest risk) and quintile 5 receives a risk score of 5 (highest risk). This data-driven approach ensures objective risk assessment based on actual payment performance rather than subjective judgments. Different customer categories

show different payment behaviors based on billing cycles, payment methods, and economic stability patterns. Residential customers generally show different risk characteristics compared to commercial enterprises or government institutions due to varying cash flow patterns and administrative processes. This feature applies operational experience through weighted risk factors, enabling the segmentation model to utilize institutional knowledge about customer type reliability rather than relying solely on demographic characteristics.

Table 1. Engineered Features for Customer Segmentation

Feature	Formula	Description	Purpose
Debt Efficiency	Outstanding Amount (IDR) ÷ Monthly Consumption (cubic meters)	How much debt per cubic meter of water used	Compares payment efficiency across different consumption levels
Payment Behavior Score	Scale payment delays from 1 to 10 using min-max scaling [17]	Converts months of arrears into a 1-10 risk score	Makes payment delays comparable across all customers
Category Risk Score	Assign risk weights to each customer type	Different customer types get different risk scores	Uses data-driven business knowledge about which customer types pay better
Geographic Risk Score	Assign accessibility scores to each location	Urban/rural and regional economic factors	Considers how location affects collection difficulty
Consumption Intensity	Natural log of monthly water consumption [18]	Log transformation of water usage	Reduces the impact of extremely high consumption customers
Financial Risk Score	(Outstanding Amount × Months of Arrears) ÷ Water Consumption Volume	Combines debt, delays, and usage into one score	Overall financial risk indicator for prioritizing collections

The Geographic Risk Score addresses variations in collection efficiency due to territorial accessibility and infrastructure constraints in utility service areas, resulting in scores ranging from 1 (lower risk areas) to 5 (higher risk areas). Similarly, geographic areas are ranked by their average debt efficiency and divided into quintiles for automatic score assignment. This method quantifies geographic factors affecting collection difficulty and resource allocation requirements based on actual payment performance data rather than geographic assumptions. Rural locations often require longer travel times for field collection activities, while urban areas may offer economies of scale but present different socioeconomic challenges that affect payment behavior. The geographic component ensures that segmentation results reflect operational realities of collection management across different service territories.

Financial Risk Score provides a composite assessment that integrates multiple risk dimensions into a unified indicator for collection prioritization, yielding values that can range from low hundreds (customers with small debts, short delays, and high consumption) to tens of thousands (customers with large outstanding amounts, extended delays, and low consumption). While individual risk metrics may present incomplete risk assessments, this multiplicative approach amplifies risk signals when multiple factors align negatively, providing an evaluation of customer financial stability. For instance, a customer with moderate debt but extended payment delays and low consumption represents a different risk profile than a customer with high absolute debt but recent payment delays and proportional consumption levels. This composite scoring mechanism enables collection teams to prioritize interventions based on multiple risk factors rather than single-dimensional assessments.

2.3. Data Preprocessing and Validation

Data preprocessing ensures data quality and algorithm compatibility through systematic validation and transformation procedures. The preprocessing pipeline addresses three key requirements: data integrity verification, outlier management, and feature standardization for clustering algorithm optimization.

Data quality assessment confirms completeness and consistency across all customer records in the dataset. Missing value analysis verifies data availability for all variables required for feature engineering calculations. Data type validation ensures numerical variables maintain appropriate formats for mathematical operations, while categorical variables provide consistent classification schemes necessary for risk score assignment.

Outlier detection employs the Interquartile Range (IQR) method to identify extreme values that may represent data entry errors or exceptional customer circumstances [17]. The IQR threshold of 1.5 times the interquartile range is a standard statistical convention for identifying outliers, balancing sensitivity with specificity in detecting genuine anomalies. This approach preserves legitimate high-risk customers while identifying potential data quality issues that could affect clustering performance.

Feature standardization applies z-score normalization to the six engineered features to ensure equal contribution to the clustering algorithm [18]. Standardization is essential for K-means clustering because the algorithm relies on Euclidean distance calculations, which features with larger numerical scales can dominate. The engineered features exhibit different scales: Debt Efficiency and Financial Risk Score produce values in thousands, Payment Behavior Score ranges from 1 to 10, Category and Geographic Risk Scores range from 1 to 5, and Consumption Intensity produces logarithmic values. The normalization process transforms all six

engineered features to have zero mean and unit variance, enabling the clustering algorithm to weight each feature equally regardless of original measurement units or scales. Raw data columns are not normalized as they are not directly used in the clustering analysis.

2.4. K-means Clustering

The clustering implementation employs scikit-learn's K-means algorithm in Python with Euclidean distance metrics and k-means++ initialization [12, 13]. Euclidean distance was selected for its suitability with standardized continuous features, while z-score normalization ensures equal feature contribution to distance calculations. The k-means++ initialization method improves convergence reliability by selecting initial cluster centroids distributed across the feature space, reducing the likelihood of poor local optima. Algorithm parameters are configured with a maximum of 300 iterations and a convergence tolerance of 1e-4 to ensure adequate optimization while maintaining computational efficiency for the 1,500-record dataset.

Hyperparameter optimization employs a systematic multi-criteria evaluation framework to determine optimal cluster number, integrating quantitative metrics with qualitative business considerations. The evaluation process examines cluster configurations from k=2 to k=6, establishing an upper bound that maintains operational feasibility for collection management implementation while providing sufficient range for identifying natural data structure.

The elbow method uses the Within-Cluster Sum of Squares (WCSS) evaluation to identify diminishing returns in variance explanation across increasing cluster numbers [18]. WCSS is calculated as the sum of squared Euclidean distances between each data point and its assigned cluster centroid, with lower values indicating tighter cluster cohesion. The elbow point represents the cluster number beyond which additional clusters provide marginal improvement in WCSS reduction relative to increased model complexity. For our analysis, WCSS values are computed across k=2 to k=6, with diminishing returns identified when WCSS reduction falls below 15% threshold, indicating limited benefit from additional cluster subdivision.

Silhouette coefficient analysis provides a complementary assessment of cluster quality by measuring both cohesion (how close each point is to others in its cluster) and separation (how far each point is from the nearest neighboring cluster) [18]. For each data point i, the silhouette coefficient s_i is calculated as in 1, where $a_{(i)}$ represents the mean distance to other points in the same cluster and $b_{(i)}$ represents the mean distance to points in the nearest neighboring cluster. Overall silhouette scores range from -1 (poor clustering) to +1 (excellent clustering), with interpretation thresholds: $s_i < 0.3$ (poor), $0.3 \leq s_i < 0.5$ (reasonable), $s_i \geq 0.5$ (good), $s \geq 0.7$ (strong).

$$s_{(i)} = \frac{b_{(i)} - a_{(i)}}{max a_{(i)}.b_{(i)}} \tag{1}$$

The optimal cluster number is determined through convergent evidence across multiple criteria [16, 18]. First, the elbow method identifies the WCSS inflection point, indicating diminishing returns. Second, silhouette analysis confirms cluster quality exceeds a reasonable threshold ($s_i \geq 0.3$). Third, statistical validation ensures significant differences between clusters (p ¡ 0.05). Fourth, business interpretability assessment confirms operational feasibility for collection management. This multi-criteria approach prevents the selection of statistically optimal but operationally impractical solutions, ensuring segmentation results provide actionable insights for utility management implementation.

2.5. Statistical Validation

Statistical validation employs independent samples t-tests to confirm differences between identified customer segments across the engineered features [19]. The analysis tests the null hypothesis that customer segments exhibit identical feature distributions, with significance levels set at $\alpha = 0.05$. Effect size calculations using Cohen's d measure the practical significance of observed differences [20] as in 2. Cohen's d is calculated as the difference between group means $(M_1 - M_2)$ divided by the pooled standard deviation (SD_{pooled}), where M_1 and M_2 represent the respective group means and SD_{pooled} represents the pooled standard deviation across both groups. Effect size interpretation follows established conventions where d = 0.2 indicates small effects, d = 0.5 indicates medium effects, and d = 0.8 indicates large effects, enabling evaluation of whether segment differences represent meaningful distinctions.

$$Cohen's = \frac{M_1 - M_2}{SD_{pooled}} \tag{2}$$

Cluster quality assessment utilizes silhouette analysis at both the dataset and individual data point levels to evaluate clustering effectiveness. Overall silhouette scores assess global clustering quality, while individual silhouette coefficients identify potential misclassified customers and cluster boundary uncertainties. The percentage of data points with positive silhouette coefficients provides a measure of clustering success and algorithm appropriateness for the dataset characteristics. Silhouette scores above 0.5 indicate good clustering structure, while scores below 0.3 suggest poor clustering quality.

3. RESULT AND ANALYSIS

This section presents the results of the K-means clustering analysis for utility customer segmentation following the methodology. The analysis progresses through feature engineering implementation, hyperparameter optimization, clustering results, and statistical validation to establish the effectiveness of the proposed segmentation framework.

3.1. Feature Engineering Results

The feature engineering process transformed raw customer billing data into six engineered features: Debt Efficiency, Payment Behavior Score, Category Risk Score, Geographic Risk Score, Consumption Intensity, and Financial Risk Score. These features are designed to capture payment patterns, consumption behaviors, and territorial risk factors. The engineered features exhibit distributional characteristics that enable clustering analysis while maintaining business interpretability. This transformation addresses the critical need for standardized risk assessment metrics that can differentiate customer payment behaviors across diverse utility customer populations.

Debt Efficiency calculations demonstrate diversity across the customer base, with values ranging from IDR 4,865 to IDR 1,028,722 per cubic meter (mean = IDR 37,064, median = IDR 28,215). This distribution indicates heterogeneity in payment efficiency relative to service consumption, validating the need for a normalized debt assessment rather than focusing on absolute outstanding amounts. The right-skewed distribution pattern is typical of financial metrics in utility customer populations, where most customers demonstrate reasonable payment efficiency while a subset exhibits higher debt-to-consumption ratios. The distribution analysis reveals that 75% of customers maintain debt efficiency below IDR 46,850 per cubic meter, indicating relatively reasonable payment behavior. However, the upper quartile demonstrates concerning payment patterns, with the top 10% of customers exhibiting debt efficiency exceeding IDR 78,000 per cubic meter. The highest debt efficiency value of IDR 1,028,722 per cubic meter represents a customer with IDR 26,315,500 outstanding debt consuming only 29 cubic meters monthly, indicating extended payment delay pattern that requires immediate intervention. Customer category analysis within debt efficiency reveals patterns aligned with business expectations. Government institutions demonstrate the highest average debt efficiency at IDR 107,279 per cubic meter, reflecting administrative delays in payment processing typical of government entities. Commercial categories show elevated debt efficiency ranging from IDR 164,430 to IDR 186,769 per cubic meter, indicating cash flow challenges in business operations. Conversely, social categories (Sosial A, MBR A) exhibit the lowest debt efficiency values below IDR 20,000 per cubic meter, reflecting the effectiveness of social assistance programs in maintaining payment compliance.

The Payment Behavior Score standardization converted payment delays, ranging from 5 to 28 months, into a 1-10 risk scale (mean = 1.72, median = 1.20). The concentration of scores toward the lower end of the scale indicates that most customers maintain relatively prompt payment patterns despite having arrears, with only a subset demonstrating extended payment delays. This distribution pattern enables the clustering algorithm to differentiate between customers with minor payment delays versus those with extended payment delay patterns. Detailed analysis of payment behavior distribution shows that 68.3% of customers score below 2.0, indicating an arrears duration of 7 months or less. The middle range (scores 2.0-5.0) comprises 28.1% of customers with moderate payment delays of 8 to 15 months. The high-risk range (scores above 5.0) represents only 3.6% of customers but includes those with arrears exceeding 16 months, indicating extended payment delays requiring intensive collection strategies.

The implementation of Category Risk Score and Geographic Risk Score through quintile-based ranking produced balanced distributions across the 1-5 risk scale. The category risk assessment identified variations in payment behavior across customer types, with commercial enterprises and government institutions generally exhibiting higher debt efficiency (higher risk) compared to residential customers and social categories. The quintile-based category scoring reveals clear differentiation in risk levels. Score 5 (highest risk) includes Large Business (186,769 IDR/m^3), Medium Business (164,430 IDR/m^3), and Provincial Government Institutions (107,279 IDR/m^3), representing 1.3%, 1.1%, and 0.4% of the customer base, respectively. Score 1 (lowest risk) encompasses Social A (10,012 IDR/m^3), Low-Income Housing A (19,392 IDR/m^3), and Social D (20,557 IDR/m^3), representing the most payment-compliant customer segments.

Geographic risk analysis through quintile ranking identified territorial risk patterns that reflect socioeconomic and infrastructure factors. High-risk geographic areas (Score 5) include Mataram Timur (68,262 IDR/m^3), Batu Layar (65,264 IDR/m^3), and Selaparang (49,142 IDR/m^3), indicating urban concentration areas with elevated payment challenges. Low-risk areas (Score 1) include Sekotong (27,541 IDR/m^3), Monjok (28,745 IDR/m^3), and Pagutan (28,836 IDR/m^3), representing stable payment territories with favorable collection conditions.

Category and Geographic Risk Scores are assigned using quintile-based ranking as a standardized scoring methodology. The observed correlation between these scores and debt efficiency ($r = 0.74$, $p < 0.001$) demonstrates internal consistency of the ranking approach rather than external validation. Similarly, geographic risk scores show correlation with territorial collection performance ($r = 0.68$, $p < 0.001$), validating the effectiveness of location-based risk assessment.

Consumption Intensity log-transformation normalized the consumption distribution from a right-skewed pattern (skewness = 4.12) to a more symmetric distribution (skewness = 0.89), reducing the impact of high-consumption outliers while preserving the behavioral patterns that distinguish different customer usage profiles. The transformation enables the clustering algorithm to focus on consumption behavior patterns rather than absolute volume differences. The log-transformed consumption values range from 2.48 to 5.70 (mean = 3.24, median = 3.18), effectively normalizing the original consumption range of 11 to 297 cubic meters. This transformation reveals that consumption patterns follow a log-normal distribution typical of utility usage, where most customers cluster around moderate consumption levels. At the same time, a few demonstrate extremely high usage patterns.

The integration of Financial Risk Score provided composite risk indicators ranging from 290 to 444,410 (mean = 9,750, median = 4,580), amplifying risk signals when multiple negative factors converge in individual customer profiles. The multiplicative nature of this score effectively identifies customers whose high debt, extended delays, and low consumption create compounded risk scenarios that require priority attention. Analysis of Financial Risk Score distribution reveals three distinct risk clusters: low-risk customers (scores below 2,000, 67.2% of population) with manageable debt and reasonable consumption; moderate-risk customers (scores 2,000-10,000, 24.8% of population) requiring monitoring; and high-risk customers (scores above 10,000, 8.0% of population) requiring immediate intervention. The highest score of 444,410 represents a customer with IDR 26,315,500 debt, 28 months of arrears, and 20 cubic meters of consumption, exemplifying the compounded risk scenario this metric identifies.

3.2. Hyperparameter Tuning and Model Selection

Hyperparameter optimization using the combined elbow method and silhouette coefficient analysis, as shown in Table 2, identified a suitable clustering configuration for the engineered feature set. The evaluation process analyzed cluster numbers from k=2 to k=6 to maintain practical utility for customer management applications. This approach aims to ensure that the selected clustering solution balances statistical optimality with operational feasibility for utility management implementation.

Table 2. Cluster Analysis Results

K	WCSS	Silhouette Score
2	**4,672.94**	**0.538**
3	3,414.72	0.408
4	2,671.72	0.363
5	2,392.38	0.321
6	2,021.71	0.326

The elbow method analysis demonstrates diminishing returns in WCSS reduction beyond k=2, with subsequent cluster additions providing marginal improvement in variance explanation. Silhouette coefficient analysis confirms k=2 as the optimal configuration with a score of 0.538, indicating good cluster separation and cohesion. This score exceeds the threshold for acceptable clustering quality ($>$0.3) and approaches the threshold for good quality ($>$0.5), demonstrating that the two-cluster solution provides meaningful differentiation between customer risk profiles. Higher cluster numbers show declining silhouette scores, suggesting that additional segments introduce subdivisions that weaken cluster cohesion. The silhouette score progression (0.538$\longrightarrow$0.408$\longrightarrow$0.363$\longrightarrow$ 0.321 $\longrightarrow$0.326) demonstrates consistent quality degradation beyond k=2, supporting the two-cluster segmentation as the most suitable.

3.3. Segment Characteristics and Statistical Validation

The optimal two-cluster solution identified customer segments with statistically different characteristics across the engineered features ($p < 0.001$). The segmentation reveals patterns in customer risk behavior that enable targeted management strategies for payment collection optimization. These customer profiles provide insights for developing differentiated intervention approaches based on observed risk characteristics.

Table 3. Customer Segment Characteristics

Metric	Moderate-Risk Segment (n=1,300, 86.7%)	High-Risk Segment (n=200, 13.3%)	t-statistic	p-value	Cohen's d
Outstanding Amount (IDR million)	0.7	3.5	15.73	$<$0.001	1.89
Debt Efficiency (IDR/m^3)	28,785	90,878	12.45	$<$0.001	1.42
Payment Behavior Score	1.26	2.06	8.92	$<$0.001	1.15
Financial Risk Score	5.4	38.3	11.34	$<$0.001	1.28
Arrears Duration (months)	6.2	9.8	18.56	$<$0.001	2.12
Consumption (m^3/month)	26.8	35.4	4.21	$<$0.001	0.52

Table 3 presents the statistical validation results demonstrating significant differences between customer segments across all measured characteristics. These statistical differences translate to practical distinctions: high-risk customers have 5x higher outstanding amounts and 58% longer arrears duration, indicating operationally meaningful segmentation with estimated 3-4x collection cost differences. These differences confirm that the segmentation framework successfully identifies customer groups requiring fundamentally different collection strategies and resource allocation approaches.

The Moderate-Risk Customer Segment comprises 1,300 customers (86.7%) characterized by manageable debt profiles and relatively stable payment patterns. This segment exhibits mean outstanding amounts of IDR 0.69 million with debt efficiency of IDR 28,785 per cubic meter, indicating reasonable payment behavior relative to service consumption. The standard deviation of IDR 0.52 million in outstanding amounts suggests relatively homogeneous debt levels within this segment. Detailed analysis reveals that 78.5% of moderate-risk customers maintain outstanding amounts below IDR 1 million, with arrears duration averaging 6.2 months. The payment behavior pattern shows 82.1% of customers scoring below 1.5 on the payment behavior scale, indicating minimal delays beyond the 4-month minimum threshold for dataset inclusion. Geographic distribution shows these customers are well-represented across all service areas, with concentrations in stable territories like Lembar (97.5%) and Sandubaya (92.1%).

The High-Risk Customer Segment includes 200 customers (13.3%) demonstrating elevated financial risk indicators across all metrics. Mean outstanding amounts of IDR 3.54 million, combined with a debt efficiency of IDR 90,878 per cubic meter, indicate payment inefficiency relative to consumption levels. The large standard deviation (IDR 2.89 million) suggests heterogeneous debt patterns within this high-risk classification. Within the high-risk segment, 34.5% of customers maintain outstanding amounts exceeding IDR 3 million, with 12.0% surpassing IDR 7 million. Arrears duration analysis reveals that 23.5% of high-risk customers exceed 12 months in payment delays, with some reaching up to 28 months. Geographic concentration analysis reveals patterns indicating elevated collection risk, with 18.4% of Cakranegara customers classified as high-risk compared to only 2.5% in Lembar.

The statistical validation confirms differences between segments across the engineered features, with consistently large effect sizes (Cohen's $d > 0.8$) indicating substantive rather than merely statistical differences. The effect sizes for outstanding amounts ($d = 1.89$) and arrears duration ($d = 2.12$) demonstrate strong differentiation between customer risk profiles. These effect sizes meet established benchmarks for operational relevance, indicating that the segmentation identifies customer groups with different risk characteristics that support distinct management approaches.

3.4. Geographic Distribution Analysis

As shown in Table 4, these customer segments demonstrate distinct geographic distribution patterns that inform territorial collection strategies. Cakranegara contains 34 high-risk customers (49% of all high-risk cases) within a single area, while other territories show lower concentrations ranging from 3 to 13 high-risk customers each. This uneven distribution suggests that collection approaches may need to vary by territory.

The concentration patterns indicate potential operational considerations for resource allocation. Areas with higher customer density per location, such as Cakranegara, may allow for more focused collection efforts with reduced travel requirements. Territories with fewer dispersed high-risk customers may require different routing approaches that balance travel costs with coverage needs. The territorial distribution provides practical information for collection planning. High-concentration areas may benefit from regular collection presence, while low-concentration areas may be suitable for combined routing strategies. These geographic patterns complement the customer risk profiles identified through segmentation, offering additional considerations for collection resource deployment across the utility's service territory.

Table 4. Geographic Distribution by Risk Segment

Sub-district	Total Customers	Moderate-Risk	High-Risk
CAKRANEGARA	185	151 (81.6%)	34 (18.4%)
SANDUBAYA	151	139 (92.1%)	12 (7.9%)
LEMBAR	120	117 (97.5%)	3 (2.5%)
LABUAPI	107	100 (93.5%)	7 (6.5%)
SEKARBELA	87	74 (85.1%)	13 (14.9%)

3.5. Cluster Quality and Validation Results

Validation analysis confirms the statistical reliability and practical significance of the clustering solution. Multiple validation approaches demonstrate that the segmentation framework produces stable, meaningful customer groupings suitable for operational implementation. The convergent validation through silhouette analysis, statistical testing, and stability assessment provides evidence for the reliability and applicability of the proposed segmentation framework.

Silhouette Analysis Results show that 89.3% of customers exhibit positive silhouette coefficients, indicating appropriate cluster assignments with good separation between segments. The overall silhouette score of 0.538 exceeds the threshold for good clustering quality (>0.5), confirming that the two-cluster solution provides differentiation between customer risk profiles. Moreover, Statistical Validation through independent samples t-tests demonstrates differences ($p < 0.001$) between customer segments across the engineered features. Effect size calculations reveal consistently large effects (Cohen's $d > 0.8$) for key financial metrics, confirming that segment differences represent meaningful business distinctions rather than statistical artifacts.

Finally, the model stability assessment, conducted through comprehensive sensitivity analysis, examined clustering robustness across multiple algorithm configurations. Across 10 independent executions with different random initialization seeds, cluster assignments demonstrated high consistency with only 1.8% average variation in customer segment allocation. The silhouette scores ranged from 0.535 to 0.542 (mean = 0.538, SD = 0.002), indicating minimal performance variation despite different starting conditions. Additional robustness testing examined modifications to convergence criteria, where tolerance levels of 1^{e-3} and 1^{e-5} produced identical cluster assignments to the standard 1e-4 setting, confirming algorithmic convergence stability. This comprehensive stability assessment indicates that the segmentation framework produces reliable, reproducible results, making it suitable for operational decision-making environments where consistency is essential for implementing collection strategies.

Additional cluster validity assessment confirms clustering robustness through complementary metrics. The Davies-Bouldin Index achieves 0.73 (a threshold of <1.0 indicates good separation), while the Calinski-Harabasz score of 1,247.3 substantially exceeds the minimum threshold of 100, both confirming a well-defined cluster structure. These indices provide convergent evidence supporting the two-cluster solution's statistical validity.

3.6. Business Implications and Strategic Recommendations

The segmentation results establish a foundation for developing targeted customer management frameworks that balance collection optimization with service quality maintenance. The distinct customer segments identified through the clustering analysis demonstrate differentiated risk characteristics, necessitating tailored operational approaches aligned with specific risk profiles and resource allocation requirements. This segmentation approach supports utility organizations in developing differentiated collection strategies based on risk-calibrated intervention protocols that may improve collection effectiveness and operational efficiency.

The moderate-risk customer segment, comprising 86.7% of the customer base, exhibits payment patterns conducive to preventive management strategies. This segment's characteristics support the implementation of data-driven intervention systems, including systematic payment reminders, flexible scheduling arrangements, and proactive communication protocols designed to maintain existing payment compliance levels. The substantial size of this segment creates opportunities for standardized operational procedures that leverage economies of scale while establishing preventive mechanisms against risk category migration.

The high-risk customer segment, which represents 13.3% of the customer population, shows financial indicators that require intensive collection interventions. The concentrated nature of this segment supports resource-intensive management approaches, including direct personal engagement, individualized payment arrangement development, and specialized case management protocols. The segment's limited size enables focused resource deployment while maintaining operational cost-effectiveness through targeted intervention strategies. This segmentation enables differentiated billing strategies: moderate-risk customers receive automated payment reminders and flexible installment options through digital channels, while high-risk customers are prioritized for field collection visits and personalized payment negotiations. Concentrating 65.4% of total outstanding debt within 13.3% of customers allows for targeted allocation of collection resources, enabling substantial cost savings by avoiding universal intensive collection approaches.

Geographic risk distribution patterns identified through the segmentation analysis support territory-based resource allocation strategies that prioritize high-concentration areas such as Cakranegara while maintaining operational coverage across stable service territories. This geographic approach facilitates optimized collection route development and resource distribution based on empirically determined risk patterns rather than demographic population density, thereby enhancing operational efficiency through evidence-based territorial management. The territorial risk mapping approach supports resource allocation strategies that can adapt to geographic risk patterns while maintaining coverage across service areas.

The segmentation framework incorporates adaptive mechanisms for accommodating new customer data and evolving payment behaviors. As additional customer records accumulate, the quintile-based scoring methodology enables dynamic recalibration of category and geographic risk assessments, ensuring that risk classifications remain current with changing operational conditions. The feature engineering approach facilitates integration of new customers into existing segments through standardized calculation procedures. At the same time, the K-means clustering algorithm supports periodic model retraining to capture emerging payment behavior patterns. This adaptability ensures that the segmentation framework maintains analytical relevance as customer portfolios expand and payment environments evolve. It supports long-term operational sustainability through continuous model refinement based on accumulating empirical evidence.

While the segmentation framework successfully identifies distinct customer risk profiles and establishes the methodological foundation for targeted collection strategies, the development of specific billing policy implementations, detailed cost-benefit calculations for resource allocation, and operational deployment procedures falls outside the scope of this methodological study. The current research focuses on demonstrating the statistical validity and practical feasibility of domain-specific customer segmentation for Indonesian water utilities. The translation of these segmentation insights into concrete billing policy frameworks, staff allocation models, and collection workflow procedures represents important directions for future applied research in utility management contexts that would require collaboration with operational management teams and regulatory stakeholders to address implementation-specific requirements and constraints.

4. CONCLUSION

This research developed and validated a K-means clustering framework for utility customer segmentation through domain-specific feature engineering and statistical validation. The study addressed gaps in existing customer segmentation methodologies by developing engineered features that capture utility-specific payment behaviors, consumption patterns, and geographic risk factors for payment collection optimization. The methodology shows how machine learning techniques can be combined with domain-specific feature engineering to provide business insights for utility operations management.

The feature engineering approach transformed heterogeneous customer billing data into standardized risk assessment metrics. The development of Debt Efficiency, Category Risk Score, Geographic Risk Score, and Financial Risk Score provided risk evaluation capabilities that address limitations of traditional demographic-based segmentation approaches. The quintile-based scoring methodology for category and geographic risk assessment established an objective, data-driven risk classification that reduces subjective bias while incorporating domain-specific operational knowledge. The clustering analysis identified two customer segments with statistically significant differences across engineered features ($p < 0.001$), with effect sizes ranging from medium to large (Cohen's d = 0.52-2.12). The moderate-risk segment (86.7% of customers) exhibits manageable debt profiles suitable for standardized preventive management strategies, while the high-risk segment (13.3%) demonstrates elevated financial risk indicators requiring collection interventions. The k=2 clustering solution achieved a silhouette score of 0.538, indicating good cluster separation and validation across multiple stability assessments. Geographic distribution analysis revealed territorial risk concentration patterns, with urban areas like Cakranegara showing 18.4% high-risk customer concentration compared to 2.5% in rural areas like Lembar. These findings support evidence-based resource allocation strategies that optimize collection effectiveness through targeted territorial approaches rather than uniform population-based distribution.

This research provides a methodological foundation for Indonesian local governments and regionally-owned water companies (PTAM) to design evidence-based, risk-calibrated billing policies that enhance the financial sustainability of public water services. The framework enables a systematic transition from manual monitoring approaches to data-driven collection management systems, supporting regulatory compliance while optimizing operational efficiency. For regional water utility governance, the methodology provides tools for implementing performance-based collection strategies aligned with standards for PTAM operations. Future research will focus on developing predictive modeling capabilities for early payment risk identification, integrating with real-time digital payment systems prevalent in Indonesian financial technology applications, analyzing the temporal dynamics of customer payment behaviors, and expanding to other utility contexts, including electricity.

5. ACKNOWLEDGEMENTS

The authors would like to express their gratitude to PT Air Minum Giri Menang (PTAM Giri Menang) for providing the operational data that made this research possible. We appreciate their cooperation and support in facilitating data access for academic research purposes.

6. DECLARATIONS

AI USAGE STATEMENT

During the preparation of this work, the author(s) used Gemini to improve the language and clarity of the manuscript. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

RESEARCH ETHICS AND DATA PROTECTION

This research complies with Indonesian Personal Data Protection Law (UU No. 27/2022) through comprehensive data anonymization procedures. All customer identifying information was removed before analysis, with geographic data aggregated to prevent individual identification. PT Air Minum Giri Menang provided institutional consent for the utilization of anonymous data for academic research purposes.

AUTHOR CONTIBUTION

A.H.A. and H.W. shared contributions in conceptualization, methodology, validation, and interpretation of data. A.H.A.'s specific contributions include dataset and data collection, feature engineering, data preprocessing and validation, statistical validation, and initial manuscript writing. H.W. was responsible for supervision and manuscript revision. I.W.A.A contributed to supervision and writing. All authors approved the final version of the manuscript.

FUNDING STATEMENT

This research uses private funds from researchers.

COMPETING INTEREST

The author declares that the entire research, analysis, and manuscript preparation process was carried out without any conflict of interest that could affect the academic and scientific integrity of this article.

REFERENCES

[1] A. Ranjan and S. Srivastava, "Customer segmentation using machine learning: A literature review," vol. 2481, no. 1, p. 020036, November,2022, https://doi.org/10.1063/5.0103946.

[2] T. Stylianou and A. Pantelidou, "A machine learning approach to consumer behavior in supermarket analytics," vol. 16, p. 100600, September,2025, https://doi.org/10.1016/j.dajour.2025.100600.

[3] R. Mathumitha, P. Rathika, and K. Manimala, "Big Data Analytics and Visualization of Residential Electricity Consumption Behavior based on Smart Meter Data," in *2022 International Conference on Breakthrough in Heuristics And Reciprocation of Advanced Technologies (BHARAT)*. IEEE, April,2022, pp. 166–171, https://doi.org/10.1109/BHARAT53139.2022.00043.

[4] Q. Wang, G. Sun, F. Lou, L. Jin, and P. Lu, "Data Analytics Enabled Power Marketing Analysis and Decision-Making Supporting System," in *2022 World Automation Congress (WAC)*. IEEE, October,2022, pp. 247–251, https://doi.org/10.23919/WAC55640.2022.9934690.

[5] Y. Li, X. Chu, D. Tian, J. Feng, and W. Mu, "Customer segmentation using K-means clustering and the adaptive particle swarm optimization algorithm," vol. 113, p. 107924, December,2021, https://doi.org/10.1016/j.asoc.2021.107924.

[6] F. P. Rachman, H. Santoso, and A. Djajadi, "Machine Learning Mini Batch K-means and Business Intelligence Utilization for Credit Card Customer Segmentation," vol. 12, no. 10, 2021, https://doi.org/10.14569/IJACSA.2021.0121024.

[7] M. Pradana, "Maximizing Strategy Improvement in Mall Customer Segmentation using K-means Clustering," vol. 2, no. 1, January,2021, https://doi.org/10.47738/jads.v2i1.18.

[8] J. Salminen, M. Mustak, M. Sufyan, and B. J. Jansen, "How can algorithms help in segmenting users and customers? A systematic review and research agenda for algorithmic customer segmentation," vol. 11, no. 4, pp. 677–692, December,2023, https://doi.org/10.1057/s41270-023-00235-5.

[9] M. Alves Gomes and T. Meisen, "A review on customer segmentation methods for personalized customer targeting in e-commerce use cases," vol. 21, no. 3, pp. 527–570, September,2023, https://doi.org/10.1007/s10257-023-00640-4.

[10] I. Daniel, N. K. Ajami, A. Castelletti, D. Savic, R. A. Stewart, and A. Cominola, "A survey of water utilities' digital transformation: Drivers, impacts, and enabling technologies," vol. 6, no. 1, p. 51, July,2023, https://doi.org/10.1038/s41545-023-00265-7.

[11] S. Veltri, M. E. Bruni, G. Iazzolino, D. Morea, and G. Baldissarro, "Do ESG factors improve utilities corporate efficiency and reduce the risk perceived by credit lending institutions? An empirical analysis," vol. 81, p. 101520, April,2023, https://doi.org/10.1016/j.jup.2023.101520.

[12] M. Gul and M. A. Rehman, "Big data: An optimized approach for cluster initialization," vol. 10, no. 1, p. 120, July,2023, https://doi.org/10.1186/s40537-023-00798-1.

[13] K. Tabianan, S. Velu, and V. Ravi, "K-Means Clustering Approach for Intelligent Customer Segmentation Using Customer Purchase Behavior Data," vol. 14, no. 12, p. 7243, June,2022, https://doi.org/10.3390/su14127243.

[14] "Undang-undang (UU) Nomor 27 Tahun 2022," Oktober,2022.

[15] A. Mumuni and F. Mumuni, "Automated data processing and feature engineering for deep learning and big data applications: A survey," vol. 3, no. 2, pp. 113–153, March,2025, https://doi.org/10.1016/j.jiixd.2024.01.002.

[16] g.-i. family=Zubair, given=Md., M. A. Iqbal, A. Shil, M. J. M. Chowdhury, M. A. Moni, and I. H. Sarker, "An Improved K-means Clustering Algorithm Towards an Efficient Data-Driven Modeling," vol. 11, no. 5, pp. 1525–1544, October,2024, https://doi.org/10.1007/s40745-022-00428-2.

[17] J. VanderPlas, "Python Data Science Handbook: Essential Tools for Working with Data," 2023.

[18] A. Geron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media, Inc., October,2022.

[19] Y. AbdulRaheem, "Statistica
l Significance versus Clinical Relevance: Key Considerations in Interpretation Medical Research Data," vol. 49, no. 6, pp. 791–795, November,2024, https://doi.org/10.4103/ijcm.ijcm_601_23.

[20] S. Panjeh, A. Nordahl-Hansen, and H. Cogo-Moreira, "Establishing new cutoffs for : application using known effect sizes from trials for improving sleep quality on composite mental health," vol. 32, no. 3, p. e1969, September,2023, https://doi.org/10.1002/mpr.1969.

[This page is intentionally left blank.]